Sleep Staging Using Wearables and Deep Neural Networks

Shaun Davidson[1], Cristian Roman[1], Jonathan Carter[1], Mirae Harford[2], and Lionel Tarassenko[1]

Abstract— There is a well-established association between sleep and health status, but the current gold-standard for analysing sleep, polysomnography, is too disruptive and expensive to enable longitudinal monitoring. There is, therefore, a growing interest in automated sleep scoring, or staging, using a combination of wearable technology to acquire cardio-respiratory vital signs and machine learning to learn how these vital signs vary with sleep state. However, sleep and the associated cardio-respiratory signals also change significantly with age, in part because of age-related changes in the autonomic nervous system, and this impacts the accuracy of wearable sleep staging methods. This paper investigates how the accuracy of a deep neural network model trained on the Sleep Heart Health Study database varies with the age of the subject. We show that the classification accuracy for each sleep stage decreases with age. We also present proof-of-concept analysis of longitudinal sleep data from a COVID-19 Challenge Study with a younger cohort (18 - 29 years of age), discuss the impact of having trained the deep neural network model on a database with an age range from 40 to 89+, and suggest how this issue may be addressed.

Clinical relevance— This paper highlights how changes in sleep behaviour with age can affect neural network sleep staging using cardio-respiratory vital signs and machine learning, resulting in less accurate sleep staging in some age groups, and discusses potential methods for addressing this.

I. INTRODUCTION

There is a well-established association between sleep and health. Individuals with chronically disrupted sleep patterns have an elevated risk not only of cardiovascular diseases such as stroke and myocardial infarction [1] but also of neurodegenerative diseases such as dementia and Parkinson's disease [2]. Further, reduction in the duration of slow wave (or deep) sleep, an important sleep state associated with memory reinforcement, immune function, and repair of tissues, has been shown to be associated both epidemiologically with type 2 diabetes and acutely with reduced glucose tolerance [3].

The gold standard for sleep analysis is polysomnography (PSG), which requires a variety of specialised equipment and contact sensors used in a sleep laboratory to record the brain's electrical activity patterns using the electroencephalogram (EEG), as well as chin, eye, and limb movement, and vital signs [4]. Once all the data has been recorded for an entire night, each 30-second 'epoch' of the PSG is reviewed by at least one trained expert who manually assigns it to one of 5 states, or 'sleep stages', by an expert: wake, N1 (light), N2 (intermediate), N3 (slow wave or deep), or REM (rapid eye movement) sleep. However, PSG is expensive, can disrupt the subject's sleep, and requires several hours of expert time for manual sleep staging [4]. Further, the demands both on the subject and sleep expert are such that longitudinal studies in a sleep laboratory using PSG are impractical, making it difficult to investigate night-to-night variations in sleep patterns.

There has, therefore, been growing interest in methods for automated sleep staging using signals that can be acquired in the home environment using wearables, with minimal cost or disruption to the individual. Actigraphy, using wrist-worn wearables, has been shown to perform well when differentiating between sleep and wakefulness, but less well when differentiating between sleep stages [5]. Cardio-respiratory vital signs, in particular heart rate (HR) and respiratory rate (RR), are associated with sleep stages via the autonomic nervous system (ANS) [6] and can be easily measured using wearable technology [7], [8] such as wrist-worn sensors (as in reflectance polyplethysmography (PPG)) or chest-worn patches. The advantage of the latter is that a direct measurement of RR is available through impedance pneumography, whereas wrist-worn sensors only provide an indirect assessment through heart rate variability. It is common for wearable sleep staging to combine N1 and N2 sleep into a single 'light sleep' stage [8], thus classifying sleep into one of 4 stages (wake, light, N3, REM).

However, autonomic nervous activity changes with increasing age, with a significant decrease in nocturnal parasympathetic activity [9], potentially weakening the correlation between cardio-respiratory vital signs and sleep stages. Sleep architecture is also known to alter with age; for example [10] reported a significant drop in N3 (slow wave) sleep duration in men between those aged 16 - 25 years (18.9%) and those aged 36 - 50 years (3.4%). Machine learning models are usually trained to perform sleep staging without considering these demographic data, limiting their ability to account for these changes.

In this paper, we investigate whether the accuracy of a deep neural network model trained using the Sleep Heart Health Study (SHHS) database [11] varies with participant age. We also provide proof-of-concept results for longitudinal wearable sleep staging on prospective

This work was supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health.

[1]Shaun Davidson, Cristian Roman, Jonathan Carter, and Lionel Tarassenko are with the Dept. of Engineering Science, University of Oxford, United Kingdom shaun.davidson@eng.ox.ac.uk

[2]Mirae Harford is with the Nuffield Department of Clinical Neuroscience, University of Oxford, United Kingdom

data from an on-going study with a younger cohort (the Oxford COVID-19 Challenge study), and discuss the impact of age differences between training and testing data in wearable sleep staging, as well as potential mitigation strategies.

II. METHODS

A. Dataset

The Sleep Heart Health Study (SHHS), a multi-centre study in the US, was designed to investigate the association between sleep-disordered breathing and cardiovascular health [11]. It recruited 6,441 participants to its first phase (SHHS1) between November 1995 and January 1998. For each participant, a full PSG was recorded, along with manually annotated sleep stages for each 30-second epoch.

In this paper we selected recordings from the SHHS1 database that did not meet the American Academy of Sleep Medicine (AASM) criteria for moderate obstructive sleep apnoea (OSA) [12]. Exclusion of participants with OSA, a condition over-represented in the SHHS database given the study's focus on sleep-disordered breathing, ensures that our dataset only contains data from participants with respiratory patterns more typical of healthy individuals.

B. Vital Sign Estimation

1) Heart Rate Estimation: HR estimation used both the electrocardiogram (ECG) and the PPG-estimated HR (for signal quality assessment) from the SHHS1 dataset. The ECG waveform was band-pass filtered between 0.5 and 3.0 Hz and peaks identified using the Pan-Tompkins algorithm [13]. HR was then estimated by counting peaks in each 5 second window, sliding by 1 second. Segments of data for which no peaks were detected, ECG-estimated HR was constant, or PPG-reported HR was 0 were taken to represent signal corruption or disconnection, and were excluded.

Once HR was estimated for an entire night's recording, the recording was excluded if either the 90^{th} percentile absolute error between the ECG- and PPG-estimated HR for the whole recording was greater than 4 beats per minute, or if the Pearson's correlation during sleep between ECG- and PPG-estimated HR was less than 0.6.

2) Respiratory Rate Estimation: RR estimation employed three different respiratory waveforms from the SHHS1 dataset: thoracic and abdominal excursions measured with inductive bands and airflow from a nasal-oral thermocouple. The signals were band-pass filtered between 0.1 Hz and 0.7 Hz and breath detection performed using the moving average curve (MAC) peak detector [14]. RR was then estimated by counting peaks in each 25-second window, sliding by 1 second. RR estimates from these three waveforms within 2 breaths/min of one another were then combined by taking the mean. If no two estimates were within 2 breaths/min, RR was set to 0, representing low signal quality. Once RR was estimated for an entire night's recording, the recording was excluded if more than 7% of the recording was set to 0.

C. Dual-Stream Model

We trained the dual-stream deep neural network developed in [15] on the selected recordings from the SHHS1 dataset. The dual-stream architecture consists of two parallel 1D ResNet blocks, the first of which is trained on short-term cardio-respiratory data (HR and RR over a 5-minute window, centred on the PSG epoch, sampled at 1 Hz) and the second of which is trained on long-term cardio-respiratory data (HR and RR over a 50-minute window, centred on the PSG epoch, sampled at 0.1 Hz). The features output from these two ResNet blocks are then combined in a multi-layer perceptron for classification. A block diagram of the model is provided in fig. 1 and further details are available in [15].

We split recordings randomly into 70% training, 20% validation, and 10% testing data. Because of the relative prevalence of the light sleep stage, we undersampled the data from epochs in this sleep stage by a factor of 2 when training. We trained the dual-stream model using cross-entropy loss and the Adam optimiser, stopping if validation loss increased.

D. The COVID-19 Challenge Study

The proof-of-concept longitudinal data used in this paper comes from the current COVID-19 Challenge Study in Oxford, United Kingdom (CUREC R80395/RE001). This study recruits young (18 - 29 years of age), sero-positive healthy participants, who have previously recovered from a COVID-19 infection. The study aims to establish the lowest dose of the virus which can take hold in 50% of people who have previously been naturally infected. Participants are monitored with a chest-worn wearable patch (as in [16]) that provides 24-hour cardio-respiratory vital signs for a period of 7 days, encompassing periods both before and after being exposed to COVID-19.

III. RESULTS

A. Participant Demographics

Of the 6,441 participants in SHHS1, 5,804 have data available. Of these, 3,127 participants were below the threshold for moderate OSA, and 2,509 of these had recordings of sufficient quality for robust vital-sign estimation. Demographics for these participants are shown in table I, with more women than men, since there is a greater prevalence of OSA in men within the SHHS1 dataset, as in the general population [12].

Table II shows the relative incidence of sleep stages with age, corresponding to results in [17] and, similarly, showing a continuous decrease in N3 sleep in men with age (at ages beyond the 36 - 50 years in [10]), but not in women.

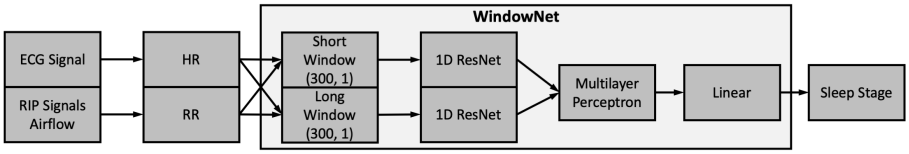

Fig. 1. Block diagram of the dual-stream model.

TABLE I
Overview of patient demographics

Demographics	Men	Women	Overall
Participants, #(%)	865 (34.5)	1644 (65.5)	2509 (100)
Age (years), mean(SD)	60.4 (11.4)	60.7 (11.3)	60.6 (11.3)
BMI, mean(SD)	27.1 (3.7)	27.0 (5.0)	27.0 (4.6)
Sleep Behaviour	Men	Women	Overall
Sleep (min), mean(SD)	360 (59)	376 (62)	371 (61)
% Wake, mean(SD)	16.2 (9.7)	14.9 (9.3)	15.4 (9.5)
% Light, mean(SD)	54.0 (10.6)	48.4 (10.8)	50.3 (11.0)
% N3, mean(SD)	12.5 (9.3)	18.9 (9.7)	16.7 (10.0)
% REM, mean(SD)	17.2 (5.8)	17.8 (5.8)	17.6 (5.8)

TABLE II
Relative incidence of sleep stages with age in the SHHS1 database. All values presented are means.

	Women				
	40 - 50	50 - 60	60 - 70	70 - 80	80+
Sleep (min)	389.6	385.1	371.4	362.1	358.1
Wake (%)	12.1	13.1	15.9	17.9	19.2
Light (%)	51.4	49.4	46.7	46.2	48.5
N3 (%)	17.0	18.8	20.0	19.6	17.6
REM (%)	19.6	18.7	17.3	16.3	14.8
	Men				
	40 - 50	50 - 60	60 - 70	70 - 80	80+
Sleep (min)	369.3	368.6	359.4	346.4	332.3
Wake (%)	13.7	14.6	16.6	19.4	23.3
Light (%)	53.4	53.7	55.4	53.4	52.9
N3 (%)	14.6	13.7	11.5	10.8	9.1
REM (%)	18.3	18.0	16.5	16.5	14.7

B. Model Performance

The training set consisted of 1,756 recordings (1,546,008 epochs), the validation set 502 recordings (443,772 epochs), and the test set 251 recordings (224,223 epochs). The model achieved an F1 score of 0.75 and Cohen's Kappa of 0.67 on the 'balanced' test set (light sleep undersampled), and an F1 score of 0.73 and Cohen's Kappa of 0.62 on the test set without undersampling. Table III shows the accuracy of the dual-stream model on the validation and test sets (without undersampling). There is a consistent decrease in accuracy with age across all sleep stages.

C. Proof of Concept

Figs. 2 and 3 show the dual-stream generated hypnograms for 30-second epochs from two consecutive nights of data for one participant in the COVID-19 Challenge study - the first night prior to exposure with the virus challenge and the second post-exposure. In each case

TABLE III
Model Accuracy on Validation and Test Sets

Validation Set						
Age Group	No.	Wake	Light	N3	REM	Overall
40 - 50	98	81.3	69.2	78.0	86.7	76.2
50 - 60	158	79.2	68.8	73.5	88.6	75.0
60 - 70	127	75.7	66.7	72.8	84.4	72.4
70 - 80	98	76.5	65.4	70.8	83.2	71.5
80+	21	72.6	60.0	64.5	64.5	64.6
Overall	502	77.6	67.3	73.3	85.8	73.4
Test Set						
Age Group	No.	Wake	Light	N3	REM	Overall
40 - 50	44	85.6	69.1	77.4	87.8	76.4
50 - 60	73	75.6	68.6	77.7	89.7	74.9
60 - 70	69	80.4	65.6	74.2	84.4	73.0
70 - 80	53	75.3	64.9	73.5	79.2	70.9
80+	12	66.6	57.2	75.0	72.9	63.7
Overall	251	77.7	66.5	75.6	85.1	73.3

the heart rate is shown below the hypnogram, and the increase in HR associated with REM sleep can be clearly seen in fig. 2. In both figures the general features of a typical sleep architecture are present (e.g., periodic REM cycles and more N3 sleep earlier in the night), with more sleep disturbance (higher amount of wakefulness) during the second night, following the challenge. We would, however, expect the N3 sleep percentage to be slightly higher in a young individuals, so it is possible that this class is being underestimated by our model.

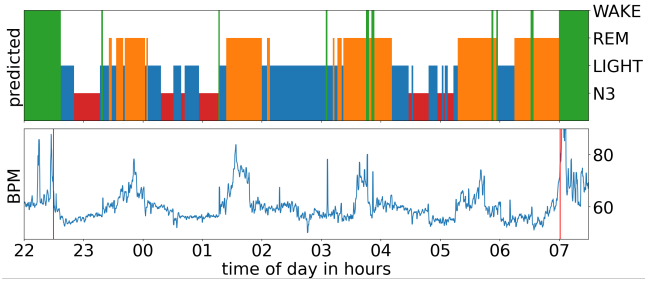

Fig. 2. Hypnogram generated by dual-stream model (top) and HR (bottom) for the first night (pre-exposure), Challenge study.

IV. DISCUSSION

The dual-stream model achieved a good overall test accuracy (73.3%) and Cohen's Kappa (0.62) on the SHHS dataset, comparable to similar methods in the literature (e.g, Cohen's Kappa of 0.65 in [8] on a cohort aged 44 - 60). Model accuracy decreased with age, a trend preserved for each sleep stage (table III). This trend is

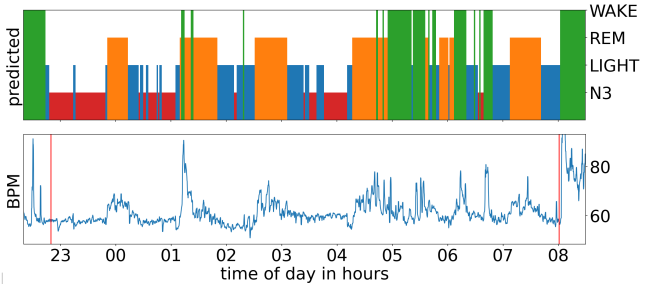

Fig. 3. Hypnogram generated by dual-stream model (top) and HR (bottom) for the second night (post-exposure), Challenge study.

likely due to a combination of the decline of ANS function and decrease in cohort size with age. Model accuracy for the 40 - 50 cohort, despite it being the second smallest cohort, is higher than that for older cohorts, emphasising the age-related influence on accuracy.

Figures 2 and 3 show the hypnograms for two consecutive nights for one of the participants in the Oxford COVID-19 Challenge study. The trained deep neural network is able to capture the sleep architecture on both nights, showing a more disturbed sleep pattern during the second night, after exposure to the virus. These proof-of-concept results demonstrate the potential of wearable, automated sleep staging using HR and RR to investigate longitudinal changes in sleep behaviour.

The dual-stream model has been trained and tested on a cohort of at least 40 years of age, with sleep-disordered breathing, in a sleep lab. The trained model is then applied to test data from a healthy participant within the 18 - 29-year age group in a COVID-19 challenge study. While our results show that the sleep staging accuracy is higher for lower age groups (table III), it is also known that there are difference in sleep architecture between the 18 - 29 and 40+ age groups, especially with regards to N3 sleep [10]. It is therefore possible that our model is underestimating the amount of N3 sleep in the test data, which would typically be slightly greater for a young, healthy volunteer. This may also be partly related to the definition of N3 sleep, which only requires the presence of 1 - 4 Hz delta waves for at least 20% of a PSG epoch.

Our dual-stream model needs to be further validated on annotated datasets containing data from healthy subjects below the age of 40 (e.g., the ISRUC-Sleep dataset [18]). There is the potential for transfer learning to be performed using these (generally smaller) datasets to better tune the model for sleep staging in younger subjects. Additionally, it may be possible to improve performance by integrating demographic information such as age directly into the network architecture.

This paper shows that longitudinal monitoring using chest-worn wearables to acquire cardio-respiratory data is feasible and yields previously unavailable information on sleep. This opens up multiple applications in healthcare, such as monitoring the onset or progression of disease, as in the COVID-19 Challenge study, and monitoring recovery following surgery or discharge from the intensive care unit, both in the hospital and at home.

References

[1] Hermida RC, Ayala DE, Portaluppi F. Circadian variation of blood pressure: the basis for the chronotherapy of hypertension. Advanced drug delivery reviews. 2007;59(9-10):904–922.

[2] Baschieri F, Cortelli P. Circadian rhythms of cardiovascular autonomic function: Physiology and clinical implications in neurodegenerative diseases. Autonomic Neuroscience. 2019;217:91–101.

[3] Tasali E, Leproult R, Ehrmann DA, Van Cauter E. Slow-wave sleep and the risk of type 2 diabetes in humans. Proceedings of the National Academy of Sciences. 2008;105(3):1044–1049.

[4] Silber MH, Ancoli-Israel S, Bonnet MH, Chokroverty S, Grigg-Damberger MM, Hirshkowitz M, et al. The visual scoring of sleep in adults. Journal of Clinical Sleep Medicine. 2007;3(02):22–22.

[5] Imtiaz SA. A systematic review of sensing technologies for wearable sleep staging. Sensors. 2021;21(5):1562.

[6] Trinder J, Kleiman J, Carrington M, Smith S, Breen S, Tan N, et al. Autonomic activity during human sleep as a function of time and sleep stage. Journal of sleep research. 2001;10(4):253–264.

[7] Pini N, Ong JL, Yilmaz G, Chee NI, Siting Z, Awasthi A, et al. An automated heart rate-based algorithm for sleep stage classification: Validation using conventional polysomnography and an innovative wearable electrocardiogram device. Frontiers in Neuroscience. 2022;16.

[8] Radha M, Fonseca P, Moreau A, Ross M, Cerny A, Anderer P, et al. A deep transfer learning approach for wearable sleep stage classification with photoplethysmography. NPJ digital medicine. 2021;4(1):135.

[9] Sato M, Betriana F, Tanioka R, Osaka K, Tanioka T, Schoenhofer S. Balance of autonomic nervous activity, exercise, and sleep status in older adults. International journal of environmental research and public health. 2021;18(24):12896.

[10] Van Cauter E, Leproult R, Plat L. Age-related changes in slow wave sleep and REM sleep and relationship with growth hormone and cortisol levels in healthy men. Jama. 2000;284(7):861–868.

[11] Newman AB, Nieto FJ, Guidry U, Lind BK, Redline S, Shahar E, et al. Relation of sleep-disordered breathing to cardiovascular disease risk factors: the Sleep Heart Health Study. American journal of epidemiology. 2001;154(1):50–59.

[12] Kapur VK, Auckley DH, Chowdhuri S, Kuhlmann DC, Mehra R, Ramar K, et al. Clinical practice guideline for diagnostic testing for adult obstructive sleep apnea: an American Academy of Sleep Medicine clinical practice guideline. Journal of Clinical Sleep Medicine. 2017;13(3):479–504.

[13] Pan J, Tompkins WJ. A real-time QRS detection algorithm. IEEE transactions on biomedical engineering. 1985;(3):230–236.

[14] Lu W, Nystrom MM, Parikh PJ, Fooshee DR, Hubenschmidt JP, Bradley JD, et al. A semi-automatic method for peak and valley detection in free-breathing respiratory waveforms. Medical physics. 2006;33(10):3634–3636.

[15] Carter J, Jorge J, Venugopal B, Gibson O, Tarassenko L. Deep Learning-Enabled Sleep Staging From Vital Signs and Activity Measured Using a Near-Infrared Video Camera. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023. p. 5939–5948.

[16] Santos MD, Roman C, Pimentel MA, Vollam S, Areia C, Young L, et al. A real-time wearable system for monitoring vital signs of COVID-19 patients in a hospital setting. Frontiers in Digital Health. 2021;3:630273.

[17] Unruh ML, Redline S, An MW, Buysse DJ, Nieto FJ, Yeh JL, et al. Subjective and objective sleep quality and aging in the sleep heart health study. Journal of the American Geriatrics Society. 2008;56(7):1218–1227.

[18] Khalighi S, Sousa T, Santos JM, Nunes U. ISRUC-Sleep: A comprehensive public dataset for sleep researchers. Computer methods and programs in biomedicine. 2016;124:180–192.